Review of MT Nutrient Criteria Documents.

Response to Questions:

MT – Wadeable Streams Draft Peer Review Questions

1. MDEQ is considering two approaches for the derivation of numeric nutrient criteria in wadeable streams: (1) eco-regional reference condition, and (2) regional and non-regional stressor-response studies. Compare and contrast the ability of each approach to provide a sound scientific basis for numeric nutrient criteria derivation. Please provide documentation on any identified ranges protective of aquatic life based on similar studies. If possible, please provide alternate methodologies using available data and tools, and describe the corresponding advantages and disadvantages.

Many of my specific concerns are detailed separately, following my responses to #1-6, some of which address this question. My overall assessment of the MDEQ dual approach was that it was generally well thought out and appeared to have protection of streams in mind. There were a few exceptions where it appeared the criteria were set a bit too high (well beyond the 100% of reference distribution), and I commented on those decisions and provided citations where available. However, MDEQ did a commendable job of reviewing scientific literature and applying peer-reviewed literature in support of developing defensible, numeric criteria.

2. In Section 3.6.1., Montana suggests that no nutrient criteria are needed for streams in the 1 Level IV Ecoregion within the Northwestern Great Plains: River Breaks (43c). The MDEQ rationale for this decision is: "This level IV ecoregion has highly turbid, flashy streams with naturally elevated TP and TN levels. Concentrations observed in the region's reference sites indicate that nutrient concentrations here are already naturally elevated above the harm-to-use thresholds identified for the plains region as a whole. As such, no nutrient criteria are recommended for streams within this level IV ecoregion." Please comment on whether the state has provided a sufficient scientific basis that 1) these levels are naturally elevated, 2) additional increase in nutrients would not cause harm to aquatic life, and 3) that, therefore, criteria are not needed. Is the reviewer aware of any additional information that could be provided to either support the State's assessment of natural background or that could be used to derive site specific criteria?

I struggled with this decision. I cannot concur that additional increases in nutrients would not cause harm to aquatic life, mainly because there are not sufficient data to support this conclusion. I commented that there needs to be some consideration of dissolved nutrients, or at least a thorough discussion about them relative to TN and TP. Potential sources of nutrients to these streams are likely to be primarily dissolved in form and, without knowing whether there are high levels of dissolved N and P in reference sites, I cannot determine whether inputs are likely to harm aquatic life. Stream flow is also an important consideration. Extended periods of low flow during droughts coupled with over-enrichment from anthropogenic sources of N and P would likely result in biological responses that could harm aquatic life. I also noted that there was a very wide range of TN and TP values among different streams, suggesting that even level 4 ecoregions may not sufficiently capture the variability in geology and natural nutrient concentrations. In sum, I suggest this decision needs further consideration.

3. MDEQ is proposing to allow TN and TP criteria to be exceeded 20% of the time and be considered supporting aquatic life uses. This frequency was derived based on analysis of the Clark Fork River chl-a data. Please comment on the proposed exceedance frequency and whether allowing the stated magnitudes to be exceeded 20% of the time would not result in adverse effects on aquatic life. This information is discussed in the State's Assessment Methodology.

Allowing TN and TP criteria to be exceeded 20% of the time brings with it uncertainty about both the timing and the magnitude of exceedance. For example, exceeding 2 months in a row, in the middle of summer when flow is low is quite different than exceeding two distinct periods during a wet year with higher than average runoff. Similarly, exceeding a criterion by an order of magnitude just one time would obviously have different implications to aquatic life than if the criterion were exceeded by a few parts per billion.

The use of the Student's t-test to compare means to the criterion is MDEQ's approach to considering magnitude of exceedance. Although I am encouraged that MDEQ recognizes magnitude and frequency as important, I am not certain the t-test method is optimal. I outline my concerns with the sampling method, samples, and analysis of data for this test under point #6, below.

4. MDEQ's criteria approach includes a Chl-a value of 125 mg/m^2 to be used as part of the related assessment information. Please comment on the selection of chlorophyll as the primary response variable, the derivation of the chlorophyll threshold, and its application as a statewide assessment indicator.

Benthic CHLA is a widely used indicator of nutrient over-enrichment so it is defensible for MDEQ to include it as a measurement endpoint. However, benthic CHLA is not a reliable indicator of nutrient overenrichment because it is highly variable temporally due to periodic sloughing/senescing, grazing (Taylor et al. 2012) and scouring by high flows. In two years of sampling wadeable streams in central Texas, we found benthic CHLA to be one of the least reliable indicators of nutrient enrichment when compared to periphyton carbon: CHLA ratios, CNP ratios, enzyme activity, primary and bacterial production, and species composition (Scott et al. 2008, King et al. 2009, Scott et al. 2009, Lang et al. 2012). The observed frequency of exceeding 125 mg/cm2 CHLA could be highly variable depending upon the natural flow regime of a stream, interannual variability in precipitation, and timing of site visits, even though a stream may be vulnerable to dense periodic blooms that result in harm to aquatic life.

I also found the use of piecewise regression models (Dodds et al. 2002) to infer chlorophyll a values at particular nutrient levels to be questionable for a few reasons. It did not appear the confidence limits were considered. The fitted mean value falls within a highly variable cloud of points, indicating that 125 mg/m2 is exceeded in many streams possibly as much as half the time. If the goal is to keep CHLA below 125 mg/m2 a certain percentage of the time, quantile regression splines (Anderson et al. 2008) or other nonlinear quantile regression method would more closely match the objective. For example, if the goal was to keep CHLA < 125 mg/m2 80% of the time, the TP or TN value that aligns with the lower 5% CI of the 20% quantile would be a more appropriate number. Thus, risk of exceeding 125 mg/cm2 seems to be potentially high, or at least high uncertain, given the approach to derived the TN/TP thresholds and the high variability in benthic CHLA during the growing season.

5. Section 4.0 outlines a process for determining reach-specific nutrient criteria. Please comment on MDEQ's proposed approach for deriving reach-specific values.

The rationale and methods for setting criteria for this reach seem defensible. The process was consistent with the process used among ecoregions. Overall it is hard to find many suggestions on how they could improve their approach for setting criteria in these rivers, however see my previous comments about using CHLA as a biological endpoint, the use piecewise regression models to identify TP and TN criteria, and several point of concern about the sample design and statistical methods (see #6, below)..

6. Montana is proposing to interpret the numeric criteria using the Students t-test and binomial test to determine whether a stream segment is impaired. Please comment on the State's rationale for this approach.

It is obvious MDEQ has given this process a great deal of thought. Overall I am encouraged by the level of detail in the process and what appears to be a sincere attempt to develop criteria and a process for assessing criteria that is protective of aquatic life in the waters of Montana. This section is particularly important because it describes the nuts-and-bolts of how criteria are used to assess compliance.

There are several moving parts in this process that have the potential to strongly influence the outcome of an assessment. First, the manner in which reaches are delineated is flexible such that it seemed a bit ambiguous to me. Because sample "sites" allocated within reaches are used to assess criteria, how reaches are delineated could be manipulated to influence the outcome of assessments.

The use of multiple sampling sites within a reach to assess criteria is reasonable, but the scale of nutrient overenrichment required to fail a reach seems to be quite large. For example, under low, summer flow conditions, one site within a reach could conceivably fail during both visits whereas downstream reaches pass each time. The use of multiple downstream sites, some of which could be many kilometers away, to calculate exceedance frequency and mean nutrient levels ignores the local impairment and effectively "dilutes" the problem at this location, despite the fact it could span > 1 mile of stream (minimum distances between sites was 1 mile, correct?).

Another factor is the manner in which sampling locations and repeated measurements from those locations are used in the binomial test and t-test as if each sample unit reflects a measurement from the same population. There are two levels of organization being mixed here. Spatial and temporal sample units are being thrown in together in a haphazard way that ignores the distinct components of variance. If there were a clearer definition of reaches, site locations, and sample frequency from those sites, I would feel a little less uneasy about it, but as it stands, I get the impression that reaches may differ wildly in length, number of sites per reach will thus differ, and sample frequency may also differ.

The Clark Fork example illustrates the problem: 15-20 individual CHLA samples were collected per date and each "sample" was treated as a repeated measure, when in fact these are subsamples that are nested within a single observational unit (a site? I can't follow the sampling design very well). The total CHLA "samples" were 285-333 per site over a multiple-year period, but there were far fewer sampling events than 285-333, and far fewer TN and TP "samples" as well because those were composite grab samples. There also were different numbers of "samples" taken per site within the Clark Fork reach, as well as different numbers of samples within a site among dates. This type of analysis would not likely hold up in a peer-reviewed journal because each CHLA measurement is subsample of a single observational unit (site). In sum, I'm not necessarily saying that the approach will lead to wildly inaccurate assessments but I do believe that there are better ways to account for multiple measurements within a site and multiple dates per site within a reach to arrive at an estimate of exceedance frequency.

I also do not really see this as a hypothesis testing problem, but rather a risk assessment or probability of exceedance problem. There is a burgeoning literature on misuse of hypothesis testing statistics for ecological risk assessment and environmental assessment. The use of this approach for this particular application does not strike me as ideal.

I also am not certain about appropriateness of a t-test to detect magnitude of exceedance relative to the criterion. The t-test is a normal-distribution statistic that will be less likely to detect a difference in the mean relative to the criterion when data are skewed, and skewed data (infrequent but large departures from the criterion) are exactly why the statistic is being computed in the first place. Several other methods could be considered, ranging from computing empirical confidence limits using the bootstrap, to more sophisticated Bayesian approaches where an appropriate sample distribution is used and the test computes the probability that the sample mean differs from the criterion.

Specific comments, Addendum:

p2-2: Equitability of sample representation. I agree this is an important consideration but do not understand how the evenness statistic was applied to address the problem. How was J computed, specifically in terms of the observations in the nutrient database? The data are nested by sample unit (site),with each observation representing a distinct date, correct? More detail is needed here.

Section 2.5.1. This paragraph is interesting and I don't have any particular problems with the content except that it does not seem to have any direct applicability to criteria development in Montana. How was the information from sites that were intentionally enriched with N and/or P used to support criteria development? The section ends by suggesting this information was valuable for establishing a "lower bounds" for nutrient concentrations, but how was the information used? What are "lower bounds"?

Section 2.6. This section is an important addition to the document. I think the idea that differential nutrient limitation among different algal and other microbial species is not sufficiently acknowledged in the development in numerical nutrient criteria. This section does an excellent job of describing why managing for 2 nutrients is critical. However, I think a couple of ideas are used interchangeably and might need to be distinguished a bit.

The most important reason for differential nutrient limitation is that different species have different relative N and P demands thus one may be predominantly limiting to an aggregate endpoint such as benthic chlorophyll but in most circumstances at least some species are limited by another resource. This appears to be particularly true of photoautotrophs and heterotrophic microbes growing together in a periphyton community (Scott et al. 2008, 2009, Lang et al. 2012). In this paragraph, the idea of different species being limited by different nutrients is introduced, but later is conflated with the idea of communities switching back and forth between N and P limitation. These are 2 distinct ideas and should be parsed as such.

It is also unclear what is meant here by limitation. Limitation of accrual of benthic chlorophyll or something else? There are numerous indicators of limitation that may not manifest themselves as an increase in standing stocks if other factors are controlling accumulation in the short run. Enzyme activities, in particular, may reveal dual limitation of different subsets of species in the community whereas total biomass remains unchanged with enrichment of N, P or both because of the decoupling of heterotroph and autotroph recycling of carbon, N and P. I say this mainly to encourage a more explicit definition of limitation and acknowledgment that biomass accumulation may not be a good indicator of limitation in all situations.

The discussion about Redfield ratios is fine to include, but again it seems to be lumping responses into one large bin of either N or P limited, when in fact differential limitation means that each species in an attached community of photoautotrophs and heterotrophs has a different N and P demand, hence a community-level N:P ratio target is naive and potentially dangerous. I think ratios are a lot less important than concentrations and supply rate (velocity). Nutrient criteria should emphasize maintaining concentrations that fall below levels of individual nutrients that are known to overstimulate algae and/or microbes; the ratios at those levels may or may not be near "Redfield" because it is the supply rate of ions to the cells that ultimately determines whether a nutrient is limiting to growth or other physiological process.

In sum, I like the fact that Montana is thinking about these details but am a bit concerned about some of the overgeneralizations about nutrient limitation and nutrient ratios in driving decisions to manage for both N and P. The decision to manage for N and P need not be any more complicated than the fact that differential limitation probably occurs in most stream ecosystems and thus both nutrients are likely to limit some facet of the community at any point in time.

Section 3.0

I like the introduction to this section, detailing how the criteria are organized and presented in the forthcoming pages.

Fig 3-1 is a nice illustration of the distribution of reference sites. I noticed here and in the 2005 document that reference sites are spatially contagious. Large areas within each ecoregion are largely unrepresented by reference locations whereas other areas have high densities of them. This is a common problem, given that human activities tend to be clumped and thus the remaining "good" places are also clumped, away from human activity. However, given that there is some mention of the need for Level IV ecoregional criteria in some Level III ecoregions, it would be helpful to know whether there are some level IV ecoregions that contain few or no reference sites.

Fig 3-2. Red dots are cities? Not all red dots are labeled.

Section 3-1. Middle Rockies

Again, noting the Redfield ratio in the criteria recommendations. I don't think there is sufficient justification for including this number given it was derived for marine phytoplankton (i.e. is the the N:P ratio of marine phytoplankton). I worry about other states focusing on this ratio as they plod forward in their development of criteria. Also, the ratios reported are based on mass not moles so if ratios are to be reported please specify that they are based on mass.

p3-3, last paragraph: The interpretation of the breakpoint regression is correct, but more specifically the level of chla/m2 has reached its maximum (the bottom is effectively covered in filamentous algae). The first section of the breakpoint regression line is a quasi-linear increase with quite a bit of scatter. I don't like the interpretation of this type of regression because in reality what is happening is that the growth rate of Cladophora is faster at higher nutrient levels but with sufficient N and P will nevertheless grow until most of the channel is covered or until a high flow event knocks it back. The problem with assuming that a certain level of N or P will keep chla/m2 below a certain level is that it assumes that on average there are sufficiently frequent spates/high flow events that will keep the growth in check. In low water years or very dry summers I highly suspect that any level of N and P that is sufficient to promote filamentous algae will lead to unacceptable levels of chla/m2 (e.g. see experimental results in King et al. 2009). . If the goal is keeping chla/m2 below a certain level, other variables (particularly frequency/timing of storm events or high flows) are needed to better estimate the likelihood of failing to meet biological criteria. As currently written, I think it is overly simplisitic.

p. 3-4 Conclusions: The section acknowledges that TP as low as 20 is associated with undesirable outcomes. The use of N:P ratio as is further used to support 30 ug/L as a TP criterion because it maintains a 10:1 NP ratio, consistent with reference streams. Are we to presume that 200 ug/L TN is also associated with undesirable biological consequences as well? The justification for using the ratio as a basis for choosing 30 ug/L instead of 20 ug/L based on biological responses is warranted here. I feel there is too much emphasis on ratios without sufficient scientific documentation of it being as or more important than concentration/supply rate by ion. I am particularly concerned about the repeated reference to Redfield ratios.

Section 3.1.1 Level IV Ecoregion within the Middle Rockies: Absaroka-Gallatin Volcanic Mountains (17ia). There are only 4 reference sites in this region. The 4 sites span a huge range of TP, with as little as 16 ug/L. I find it hard to find support for a numerical criterion that would allow a stream with 16 ug/L TP to increase to 105 ug/L TP. I am confident there would be biological consequences. How realistic would it be to set basin-specific criteria for this subregion, given that it is relatively small?

Another concern is the selection of 250 ug/L TN despite the fact that this exceeds the highest reference site by almost 100 ug/L. It seems that given the high levels of TP that are naturally available in many of these streams, that any, small input of N could lead to nuisance growth of algae. In this region, it would be helpful to know the dissolved N levels because I suspect that most of the TN is particulate.. An addition of +100 ug/L NH4-N or NO3-N could lead to a substantial biological response.

3.2. Northern Rockies: Comments re: section 3.1 apply here as well.

3.3 Canadian Rockies: The very tight, extremely low TP values among all but one of the "reference" sites suggest that the selection of 25 ug/L TP for a criterion is too. high. It is far beyond the 75th percentile of reference as well as above the 20 ug/L TP number identified by other stressor response studies from the region. Again, I struggle with the use of explanatory models for predicting mean chla/m2. In most situations if TP is elevated it will be elevated by phosphate; adding 15+ ug/L TP above the highest reference sites has a high risk of impairing streams.

3.4. Idaho Batholith. Similar thoughts—TP is < 20 across all samples in reference sites. The literature review and discussion of previous results provides reasonable support for 30 ug/L TP, but not defensibly so. Setting the criterion at 30 ug/L seems to leave the door open for a minimum of 50% increase in P loads to these streams. Given this is far beyond the 75th and 90% reference site quantiles, I think greater justification is needed, especially considering the previous ecoregion was set at 25 ug/L TP despite similar reference distributions for TP.

Section 3.5. More detail on the sources of TN and TP in this region would be helpful. Is alder or another nitrogen fixing plant abundant in the uplands here? We see high natural concentrations of N in high alder streams in glaciated portions of Alaska but very low N when alder is low (Shaftel et al. 2012). As for P, is the source volcanic? What explains the high P levels in reference sites?

Also note that the discussion justifying the choice of criteria is long and somewhat speculative, although I appreciate the level of detail.

Section 3.6. The nutrient dosing study seems like it was not used directly supporting numerical criteria in this region beyond demonstrating that dissolved N and P additions stimulate algae. The amount of dissolved nutrients added was not particularly great despite the large algal and DO response, so it concerns me to see such high recommended levels of TN and TP for the region. However, the distribution of values among reference sites does support the recommended levels, assuming the reference sites are indeed representative of streams with minimal anthropogenic nutrient inputs. The large range of values among reference sites suggests that level IV ecoregions may be needed to parse out natural variability or there are streams that probably shouldn't be considered reference sites.

Section 3.6.1. River breaks. I follow the rationale for concluding that no criteria are necessary for this region. The lack of dissolved nutrient information makes it difficult to know whether all of the nutrients, particularly P, are bound to sediment or whether there is abundant dissolved N and P. I agree that dissolved nutrients can be variable due to biological uptake but in these systems I would suggest considering dissolved N and P. Without any criteria, it still seems these streams could be vulnerable to animal waste discharges, future wastewater discharges, or other sources likely to contribute very high levels of dissolved nutrients as well as organic matter. The streams are reportedly flashy, but this suggests there are periods of extended low flows between periodic flood events that permit blooms of phytoplankton and/or shallow wate r attached algae/plants. As explained in this document, I don't think the state has presented sufficient justification for electing not to set criteria for these streams.

Section 4. Reach specific criteria.

4.1 Flint Creek. The rationale and methods for setting criteria for this reach seem defensible.

4.2 Bozeman Creek et al. Overall it is hard to find many suggestions on how they could improve their approach for setting criteria in these rivers. The approach used is consistent with how it was done among ecoregions.

References

Anderson, MJ. 2008 Animal-sediment relationships re-visited: Characterising species' distributions along an environmental gradient using canonical analysis and quantile regression splines. Journal of Experimental Marine Biology and Ecology 366, 16–27

Dodds, Walter K., V. H. Smith, and K. Lohman. 2002. Nitrogen and Phosphorus Relationships to Benthic Algal Biomass in Temperate Streams. Canadian Journal of Fisheries and Aquatic Sciences. 59: 865-874.

King, RS, JM Taylor, JA Back, BA Fulton, BW Brooks. 2009. Linking observational and experimental approaches for the development of regional nutrient criteria for wadeable streams. Final Report. CP-966137-01. US Environmental Protection Agency Region 6, 151pp. http://www.baylor.edu/content/services/document.php/95606.pdf

Lang, D.A.*, R.S. King, and J.T. Scott. 2012. Divergent responses of biomass and enzyme activities suggest differential nutrient limitation in stream periphyton. Freshwater Science 31(4):in press.

Scott, J. T.*, D. A. Lang*, R. S. King, and R. D. Doyle. 2009. Nitrogen fixation and phosphatase activity in periphyton growing on nutrient diffusing substrata: Evidence for differential nutrient limitation in stream benthos. Journal of the North American Benthological Society 28:57-68

Scott. J. T.*, J. A. Back*, J. M. Taylor*, and R. S. King. 2008. Does nutrient enrichment decouple algal-bacterial production in periphyton? Journal of the North American Benthological Society 27:332-334.

Shaftel, R. S.*, R. S. King, and J. A. Back*. 2012. Alder cover drives nitrogen availability in Kenai Peninsula headwater streams, Alaska. Biogeochemistry 107:135-148

Taylor, JM, JA Back, RS King. 2012. Grazing minnows increase benthic autotrophy and enhance the response of periphyton elemental composition to experimental phosphorus additions. Freshwater Science 31 (2), 451-462